REMARKS/ARGUMENTS

As requested by the Examiner attached is a copy of page 4 of the specification. Applicants apologize for this inadvertent clerical error.

Also attached is a copy of pages 373-383 of "Applied Mulitvariate Analysis", cited on page 15 of the specification.

The Examiner has rejected claims 1, 2, 7, 12, 13, and 15-20 as anticipated, 35 USC 102(e), by U.S. Patent Application Publication 2002/008766 by Huffman et al (Huffman), claims 3, 8, 21 and 22 as unpatentable, 35 USC 103(a), over Huffman in view of Schuba et al patent 6,724,733, and claims.

In response thereto applicants are submitting a Rule 131 Declaration by co-inventor Ricardo V. Martija establishing the conception and successful reduction to practice of applicants' invention, as described in this specification and claimed herein, prior to December 19, 2000, the effective date of Huffman. Accordingly, reconsideration and allowance of claims 1 through 22 and allowance of this application are respectfully requested. However, if the Examiner deems it would in any way expedite the prosecution of this application, he is invited to telephone applicants' attorney at the number set forth below.

A petition for a one month extension of time is enclosed.

Respectfully submitted,

R. V. Martija et al

By
James W. Falk
Attorney for Applicants
Reg. No. 16154
(732) 699-4465

APPLICATION NUMBER 09/774,976

ATTORNEY DOCKET APP 1208

PAGE 4 OF SPECIFICATION

The description of the invention and the following description for carrying out the best mode of the invention should not restrict the scope of the claimed invention. Both provide examples and explanations to enable others to practice the invention. The accompanying drawings, which form part of the description for carrying out the best mode of the invention, show several embodiments of the invention, and together with the description, explain the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

In the Figures:

Figure 1 is a block diagram of a network that includes a host locator and a plurality of monitoring stations for determining geographical locations of hosts in the network, in accordance with methods and systems consistent with the present invention;

Figure 2 is a block diagram of a host locator, in accordance with methods and systems consistent with the present invention;

Figure 3 is a block diagram of a monitoring station, in accordance with methods and systems consistent with the present invention;

Figure 4 is a flowchart of the steps performed by one or more monitoring stations for determining information about hosts in a network, in accordance with methods and systems consistent with the present invention; and

Figure 5 is a flowchart of the steps performed by a host locator for determining the geographical region of a host in a network based on sample hosts information determined by one or more monitoring stations in the network, in accordance with methods and systems consistent with the present invention.

BEST MODE FOR CARRYING OUT THE INVENTION

Reference will now be made in detail to the preferred embodiments of the invention, examples of which are illustrated in the accompanying drawings. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts.

In accordance with an embodiment of the invention, a host locator and a plurality of monitoring stations are provided to determine the geographical regions of one or more hosts in a network. The monitoring stations may be placed at different points in the network to get a broad cross-section of information about hosts in the network. A plurality of sample hosts in the network are preselected such that the

APPLICATION NUMBER 09/774,976

ATTORNEY DOCKET APP 1208

REFERENCE

"APPLIED MULTIVARIATE ANALYSIS'

pp. 373-383

R_1 excludes those z's for which the reverse inequality holds.[1] The
t is summarized below.

rem (13.2.1): Suppose z: $p \times 1$ is an observation either from a popu-
n π_1 with density $f_1(z|\Theta_1)$, or from a population π_2 with density
$\Theta_2)$, with prior probabilities for π_j of p_j, and costs of misclassifi-
n c_{ij}. If Θ_1 and Θ_2 are known, the minimum risk rule is to classify
o π_1 if

$$\frac{f_1(z|\Theta_1)}{f_2(z|\Theta_2)} \geq \left(\frac{c_{12}}{c_{21}}\right)\left(\frac{p_2}{p_1}\right) \equiv \text{constant}; \tag{13.2.1}$$

rwise classify z into π_2.
Remark: Since the populations are assumed to be continuous, equality
13.2.1) occurs with probability zero.]

.2 Classification into Many Populations

f $K > 2$, the minimum risk decision rule for known parameter ma-
es $\Theta_1, \ldots, \Theta_K$ is to classify z into π_k if

$$\sum_{\substack{i=1\\(i\neq k)}}^{K} p_i c_{ki} f_i(z|\Theta_i) < \sum_{\substack{i=1\\(i\neq j)}}^{K} p_i c_{ji} f_i(z|\Theta_i), \tag{13.2.2}$$

all $j = 1, 2, \ldots, K$; $j \neq k$. The extension of (13.2.1) to (13.2.2) is
ct and straightforward to prove.

mple (13.2.1)—Two Normal Populations: Suppose $\pi_j = N(\theta_j, \Sigma_j)$,
> 0, where $(\theta_j, \Sigma_j) \equiv \Theta_j$ is known, $j = 1, 2$. Then, from (13.2.1),
sify z into $N(\theta_1, \Sigma_1)$ if

$$\frac{|\Sigma_1|^{-1/2} \exp\{-\frac{1}{2}(z-\theta_1)'\Sigma_1^{-1}(z-\theta_1)\}}{|\Sigma_2|^{-1/2} \exp\{-\frac{1}{2}(z-\theta_2)'\Sigma_2^{-1}(z-\theta_2)\}} \geq \left(\frac{c_{12}}{c_{21}}\right)\left(\frac{p_2}{p_1}\right).$$

uivalently, classify z into π_1 if

$$- \theta_2)'\Sigma_2^{-1}(z-\theta_2) - (z-\theta_1)'\Sigma_1^{-1}(z-\theta_1)] \geq 2 \log \left[\frac{|\Sigma_1|^{1/2}}{|\Sigma_2|^{1/2}} \frac{p_2}{p_1} \frac{c_{12}}{c_{21}}\right]. \tag{13.2.3}$$

For illustration, suppose the misclassification costs are equal, the
ior probabilities are equal, and the covariance matrices are equal
$_1 = \Sigma_2 = \Sigma$) so that the populations differ only in location. The

[1] This result follows from the Neyman Pearson lemma (see, for instance, Kendall and Stuart, 1966).

right-hand side of the last inequality becomes zero and the left-hand side simplifies, to give the rule: Classify z into $N(\theta_1, \Sigma)$ if

$$[(\theta_1 - \theta_2)'\Sigma^{-1}]z \geq [\tfrac{1}{2}(\theta_1'\Sigma^{-1}\theta_1 - \theta_2'\Sigma^{-1}\theta_2)]. \tag{13.2.4}$$

Since the parameters are all known, both pairs of brackets in (13.2.4) can be computed and the inequality tested. Note that the discriminant function in (13.2.4) is linear in z in that it is of the form $a'z \geq b$, where a is a known vector and b is a known scalar. A related quantity is $D^2 \equiv (\theta_1 - \theta_2)'\Sigma^{-1}(\theta_1 - \theta_2)$; D is known as the Mahalanobis distance between the two populations.

Example (13.2.2)—Three (or More) Normal Populations: Suppose $K = 3$, and all parameters are known, $j = 1, 2, 3$. The rule in (13.2.2) becomes: Classify z into π_1 if

$$p_1 c_{21} f_1(z|\Theta_1) + p_3 c_{23} f_3(z|\Theta_3) < p_2 c_{12} f_2(z|\Theta_2) + p_3 c_{13} f_3(z|\Theta_3),$$

and

$$p_1 c_{31} f_1(z|\Theta_1) + p_2 c_{32} f_2(z|\Theta_2) < p_1 c_{21} f_1(z|\Theta_1) + p_2 c_{23} f_3(z|\Theta_3).$$

A similar result is found for classifying z into π_1 and π_2 by permuting the subscripts.

Now consider the special case in which all misclassification costs c_{ij} are equal ($i \neq j$). Then the last two equations reduce to: Classify z into π_3 if

$$p_1 f_1(z|\Theta_1) < p_3 f_3(z|\Theta_3),$$

and (13.2.5)

$$p_2 f_2(z|\Theta_2) < p_3 f_3(z|\Theta_3).$$

Since the posterior probability density that $z \epsilon \pi_j$, for given z, is proportional to $p_j f_j(z|\Theta_j)$, this result shows that when the misclassification costs are equal and the population parameters are known, the classification rule becomes: Choose that π_j which maximizes the posterior probability density associated with π_j.

The Bayesian approach toward classification when all parameters are known and misclassification costs are equal, would begin with an evaluation of the posterior probability that $z \epsilon \pi_j$ given z, for each $j = 1, \ldots, K$. Then posterior odds might be computed for each pair of populations; alternatively, with $K > 2$, the population with the greatest posterior probability density can be selected.

When the costs of misclassification are unequal, the Bayesian would select the population that produced a minimum cost when averaged with respect to the posterior distribution. But this is equivalent to the sampling theory result obtained above.

Thus, the Bayesian and sampling theory approaches lead to the same

ssification rule when the parameters are known. Moreover, this result clearly valid for all $K \geq 2$.

Now suppose the three populations are multivariate Normal; that is, $= N(\theta_j, \Sigma_j)$, $\Sigma_j > 0$, $j = 1, 2, 3$, and all parameters are known. More-er, suppose all prior probabilities are equal, and all misclassification sts are equal. Then the rule becomes: Classify $\mathbf{z}$ into π_3 if

$$f_3(\mathbf{z}|\theta_3,\Sigma_3) > f_1(\mathbf{z}|\theta_1,\Sigma_1),$$

id

$$f_3(\mathbf{z}|\theta_3,\Sigma_3) > f_2(\mathbf{z}|\theta_2,\Sigma_2);$$

at is, classify $\mathbf{z}$ into π_3 if

$$(\mathbf{z} - \theta_j)'\Sigma_j^{-1}(\mathbf{z} - \theta_j) - (\mathbf{z} - \theta_3)'\Sigma_3^{-1}(\mathbf{z} - \theta_3) > \log \frac{|\Sigma_3|}{|\Sigma_j|}, \quad (13.2.6)$$

r $j = 1$ and $j = 2$. The subscripts would merely be permuted for classifi-tion into π_1 and π_2.

For classification of $\mathbf{z}$ into one of K known multivariate Normal popu-tions, the rule is merely to classify $\mathbf{z}$ into the population with the largest nsity (for the case of equal p_j's and equal c_{ij}'s). Thus, classify $\mathbf{z}$ into if

$$(\mathbf{z} - \theta_j)'\Sigma_j^{-1}(\mathbf{z} - \theta_j) - (\mathbf{z} - \theta_K)'\Sigma_K^{-1}(\mathbf{z} - \theta_K) > \log \frac{|\Sigma_K|}{|\Sigma_j|}, \quad (13.2.7)$$

r every $j = 1, 2, \ldots, K - 1$.

[*Remark:* Note that the discriminant functions given in (13.2.6) and 3.2.7) are quadratic in $\mathbf{z}$. However, if it may be assumed that the co-riance matrices are equal, the discriminant functions become linear.]

3.3 UNKNOWN PARAMETERS

In this section the population parameters are assumed to be unknown, is usually the case. Classification procedures are developed first for servations obtained from arbitrary, continuous, multivariate distribu-ons and then for observations assumed to follow multivariate Normal stributions.

3.3.1 Arbitrary Distributions

Suppose π_j has an associated density $f_j(\mathbf{z}|\Theta_j)$, $j = 1, 2, \ldots, K$, where $_j$ is unknown. Suppose further that independent p-variate observations $\mathbf{x}_1(j), \ldots, \mathbf{x}_{N_j}(j)\}$ are available for each population j, $j = 1, \ldots, K$. Then, these samples are used to form maximum likelihood estimates of the population parameters, and if the estimates are substituted for the param-eter values in (13.2.2), large sample classification rules will be obtained.

If sample sizes are sufficiently large, results obtained by this technique should be quite good. However, with moderate or small samples, the results could be quite poor.

13.3.2 Normally Distributed Observations

Sampling Theory Background

Now suppose that $\pi_j = N(\theta_j, \Sigma_j)$, $j = 1, \ldots, K$, and (θ_j, Σ_j) are unknown. Also suppose that the independent p-variate observations from each population $\{\mathbf{x}_1(j), \ldots, \mathbf{x}_{N_j}(j)\}$ are available, $j = 1, \ldots, K$. If $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K$, likelihood ratio and similar procedures may be found easily although the distributions required to use these procedures in small samples are quite complicated (see, for instance, Wald, 1944; Anderson, 1951; and Sitgreaves, 1952). Asymptotic results were given for the general case of unequal means and unequal covariance matrices by Press (1964). A large variety of other techniques have been suggested from the sampling theory viewpoint; none of them are very simple. However, the Bayesian approach provides a useful and simple alternative in this case.

Bayesian Approach

Bayesian approaches to the classification problem in the case of Normally distributed observations with unknown parameters were discussed by Geisser (1964; 1966; and 1967) and Dunsmore (1966). The results are extremely simple to apply and there is no complicated distribution theory. The results are summarized below.

Define the sample mean and covariance matrix (unbiased estimator) for the jth population as

$$\bar{\mathbf{x}}(j) = \frac{1}{N_j}\sum_{i=1}^{N_j} \mathbf{x}_i(j), \qquad \mathbf{S}_j = \frac{1}{N_j - 1}\sum_{i=1}^{N_j} [\mathbf{x}_i(j) - \bar{\mathbf{x}}(j)][\mathbf{x}_i(j) - \bar{\mathbf{x}}(j)]',$$

and recall that p_j is the prior probability of classifying $\mathbf{z}$ into π_j, $j = 1, \ldots, K$.

Theorem (13.3.1): Let $\mathbf{z}: p \times 1$ be an observation from one of the populations $\pi_j = N(\theta_j, \Sigma_j)$, $j = 1, \ldots, K$ where the parameters (θ_j, Σ_j) are unknown. If the prior distribution of the parameters is diffuse, the predictive probability density (see Section 3.7) for classifying $\mathbf{z}$ into π_j is given by the multivariate Student t-density

$$p(\mathbf{z}|\text{data}, j) = \frac{k_j}{\left[1 + \frac{N_j}{N_j^2 - 1}(\mathbf{z} - \bar{\mathbf{x}}(j))'\mathbf{S}_j^{-1}(\mathbf{z} - \bar{\mathbf{x}}(j))\right]^{N_j/2}}, \quad (13.3.1)$$

k_j is a constant not depending upon z, and given by

$$k_j = \left[\frac{N_j}{(N_j+1)\pi}\right]^{p/2} \frac{\Gamma\left(\frac{N_j}{2}\right)p_j}{\Gamma\left(\frac{N_j-p}{2}\right)|(N_j-1)\mathbf{S}_j|^{1/2}}.$$

: The proof of this theorem is given in Complement 13.1.

mark (1): From (13.3.1) it follows that the *predictive odds ratio* lassifying z into π_i, as compared with π_j, becomes the ratio of the iated multivariate Student t-densities

$$\frac{p(z|\text{data},i)}{p(z|\text{data},j)} = L_{ij} \frac{\left\{1 + \frac{N_j}{N_j^2-1}(z-\bar{x}(j))'\mathbf{S}_j^{-1}(z-\bar{x}(j))\right\}^{N_j/2}}{\left\{1 + \frac{N_i}{N_i^2-1}(z-\bar{x}(i))'\mathbf{S}_i^{-1}(z-\bar{x}(i))\right\}^{N_i/2}}, \quad (13.3.2)$$

e L_{ij} is a constant given by

$$= \left(\frac{p_i}{p_j}\right)\left(\frac{|(N_j-1)\mathbf{S}_j|}{|(N_i-1)\mathbf{S}_i|}\right)^{1/2} \left[\frac{\Gamma\left(\frac{N_i}{2}\right)\Gamma\left(\frac{N_j-p}{2}\right)}{\Gamma\left(\frac{N_j}{2}\right)\Gamma\left(\frac{N_i-p}{2}\right)}\right] \left[\frac{N_i(N_j+1)}{N_j(N_i+1)}\right]^{p/2},$$

= 1, 2, . . . , K.]

emark (2): It is seen that the sample sizes need not be large for 3.1) to be applicable, as was the case in the sampling theory approach ection (13.3.1). Therefore, the result is more generally applicable. eover, the covariance matrices need not be equal, which is usually ired in sampling approaches.]

Remark (3): If a natural conjugate rather than a vague prior is used, it raightforward to check that the result of (13.3.2) is again obtained, pt that in this case, the location, scale, and degrees of freedom param- s are different.]

mple (13.3.1): The question of how similar are the audiences of two nore advertising vehicles may be answered, in part, by an appeal to rimination analysis. Massy (1965) used this approach to evaluate the ilarities among the audiences of 5 FM radio stations in the Boston tropolitan area. The data were collected from a sample of families who owned at least one FM radio receiver, and a mail questionnaire was used to obtain information on current station selections and some 47 socio- economic and consumption variables. Respondents were given a series of scales simulating the markings on a typical FM dial and asked to note

the position of the dial on each FM receiver in the home, as of the time the questionnaire was filled out. The result of this approach was a sample of 239 families for whom the station tuned to at response time could be unambiguously determined. This sample was used to estimate the parameters of the distributions (as in Section 13.3.1) for each population (all populations were assumed to be multivariate Normal).

Since the dimension of this problem was very large ($p = 47$), the data were first subjected to a factor analysis (see Chapter 10). This resulted in a set of 12 new variates which were used as summary or index variables for the original set. (Clearly some information was thereby lost.) The 12 remaining variables were then used in a 5-population discrimination analysis to establish a basis for classifying survey respondents into listeners of one of the 5 radio stations on the basis of their socioeconomic characteristics. The analysis resulted in a profile of socioeconomic characteristics of the listeners to each of the FM radio stations.

To carry out the analysis, Equation (13.2.7) was used assuming the a priori probabilities for each of the five populations were equal, and assuming the costs of misclassification were equal. Moreover, the densities were all taken to be Normal densities with equal covariance matrices and with parameters equal to the estimated sample values (as explained in Section 13.3.1). The 12 classification variables and their multiplying coefficients used to form the linear discriminant functions for the 5 FM stations are given in Table 13.3.1. This table was interpreted by Massy to provide the following audience profiles for each of the stations.

Station A: Ownership of a bigger or newer car, or more than one car, contributes most strongly to classification in A's audience. Families that seldom "go out" to movies, sports, or cultural events also are disproportionately likely to be A's. The younger the family, the higher its probability of being in the A audience.

Station B: The probability of classification in B increases as the family rises in occupational status. It is highest if the family did not send in for A's program guide. Younger families, and families that indicate a preference for opera over jazz, are more likely to be assigned to B.

Station C: Respondents assigned to C tend to be much older than average, and own fewer and/or older and smaller automobiles, and prefer jazz and popular music to opera. "Going out" contributes more to C's classification probability than to any other station. The same is true for sending in for A's program guide.

Station D: Individualism contributes most strongly to the probability of classification in this audience. Next in importance is occupational status.

Table 13.3.1 Discriminant Function Coefficients

Variables	Stations				
	A	B	C	D	E
Durables ownership (high scorers more likely to own dishwashers, freezers, washers, dryers, second cars)	−.18	−.53	+.27	−.74	−1.01
Age—older (+)	−.89	−.79	+1.18	+.88	−1.72
Social class I—higher occupational status	+.41	+.90	+.21	+1.22	+1.11
Musical preference I—classical and opera (+) versus popular (−)	+.03	+.06	−.26	+.11	−.01
Social class II—"lower middle class" (high scorers use credit, have low income and assets, tend to have older cars)	+.20	+.29	−.34	+.57	+.94
Automobile ownership (high scorers own newer cars, tend toward foreign, lower priced, and larger models)	+.96	−.01	−1.27	−1.04	+1.10
Music preference II—folk (+) versus popular (−)	−.04	+.19	+.06	+.18	−.27
Source of entertainment (high scorers seldom "go out")	+.58	+.21	−.48	+.24	+1.04
Wife's status—working wife (+)	+.36	−.09	−.06	−.69	−.49
Music preference III—opera (+) versus jazz (−)	+.28	+.35	−.55	−.13	+.77
"Individualism" (high scorers tend to like folk music, dislike trading stamps, and not own TV set or shop in discount houses)	−.27	+.20	+.71	+1.98	+.20
Program guide—sent in for Station A's guide (−)	+.31	+.43	+.18	+.25	+.22
ɪnstant	−.26	−.19	−.38	−.38	−.33

fluence in automobile ownership strongly inhibits the chances of being classified.

ation E: High classification probabilities for E are strongly related to occupational status and automobile affluence, and inversely related to "going out" and durables ownership. Younger people are much more likely to be classified in this audience. The group is most likely to exhibit "lower middle class" values (Social Class II). The extreme positive coefficient for opera versus jazz might be regarded as a dislike for jazz.

On the basis of the above audience profiles determined from the classification analysis, a potential advertiser would find it easy to select a particular FM station for advertising his product if he could establish the "type" of individual most likely to buy his product. Thus, if buyers of his product were given questionnaires to determine their socioeconomic backgrounds, all buyers could be classified by the above linear discriminant functions to see if most buyers would be likely to be drawn toward a particular radio station, and therefore, most likely to hear the commercial message.

Note that there has been no discussion of the validity of the assumptions of multivariate Normality, equality of covariance matrices, use of large sample results of parameter estimation, and so on. Violation of any of these assumptions would vitiate the results described above.

Example (13.3.2): Suppose $\mathbf{z}$ has two components ($p = 2$), and $\mathbf{z}$ is to be classified into one of two Normal populations ($K = 2$). Assume there is equal likelihood, a priori, of classifying $\mathbf{z}$ into the 2 populations so that $p_1 = p_2 = \frac{1}{2}$. A full Bayesian approach will be used for the classification.

Suppose that on the basis of 10 bivariate observations from each population ($N_1 = N_2 = N = 10$), the sufficient statistics are

$$\bar{\mathbf{x}}(1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \bar{\mathbf{x}}(2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\mathbf{S}_1 = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}, \qquad \mathbf{S}_2 = \begin{pmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

Since the sample sizes are equal and the prior probabilities are equal, from (13.3.2)

$$L_{ij} = \frac{|\mathbf{S}_j|^{1/2}}{|\mathbf{S}_i|^{1/2}}.$$

Since $|\mathbf{S}_1| = \frac{3}{4}$, and $|\mathbf{S}_2| = \frac{7}{16}$, $L_{12} = .76$. From (13.3.2),

$$\frac{p(\mathbf{z}|\text{data}, i = 1)}{p(\mathbf{z}|\text{data}, i = 2)} = \frac{.76\{1 + \frac{10}{99}(\mathbf{z} - \bar{\mathbf{x}}(2))'\mathbf{S}_2^{-1}(\mathbf{z} - \bar{\mathbf{x}}(2))\}^6}{\{1 + \frac{10}{99}(\mathbf{z} - \bar{\mathbf{x}}(1))'\mathbf{S}_1^{-1}(\mathbf{z} - \bar{\mathbf{x}}(1))\}^6}.$$

Suppose the observed vector to be classified is given by $\mathbf{z} = (\frac{1}{2}, \frac{1}{2})'$. Then

$$\mathbf{z} - \bar{\mathbf{x}}(1) = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \qquad \mathbf{z} - \bar{\mathbf{x}}(2) = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix},$$

$$S_1^{-1} = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}, \qquad S_2^{-1} = \begin{pmatrix} \frac{8}{7} & -\frac{4}{7} \\ -\frac{4}{7} & \frac{16}{7} \end{pmatrix}.$$

ce,

$$[z - \bar{x}(1)]' S_1^{-1} [z - \bar{x}(1)] = \tfrac{1}{3},$$
$$[z - \bar{x}(2)]' S_2^{-1} [z - \bar{x}(2)] = \tfrac{4}{7}.$$

stitution of these results into the ratio of densities gives for the pre-tive odds ratio,

$$\frac{p(z|\text{data}, i = 1)}{p(z|\text{data}, i = 2)} = .92.$$

at is, the predictive odds are slightly in favor of π_2, but not much re than the *prior odds ratio* of 1:1. This result was to be expected light of the observed data. That is, since the sample variances are so ge relative to the distances between the sample means, and since the or information sheds little additional light, the "boundary" between two populations remains quite blurred.

It might be of interest to see what would have been the result of ap-ing the sampling theory procedure used in Example (13.3.1) to this ample. Since there are two populations, (13.2.3) would be applied if parameters were all known, or if the sample sizes were very large. this case, the sample sizes are each 10, so that asymptotic theory ould not be expected to be applicable. However, what happens if the proach is used regardless?

Letting $p_1 = p_2$, and $c_{12} = c_{21}$, and substituting into (13.2.3) gives the le: Classify z into π_1 if

$$(z - \theta_2)'\Sigma_2^{-1}(z - \theta_2) - (z - \theta_1)'\Sigma_1^{-1}(z - \theta_1) \geq \log \frac{|\Sigma_1|}{|\Sigma_2|}.$$

ow replace $[\theta_1, \theta_2, \Sigma_1, \Sigma_2]$ by their sample estimates $[\bar{x}(1), \bar{x}(2), S_1, S_2]$, ven above. Then classify z into π_1 if

$$[z - \bar{x}(2)]' S_2^{-1} [z - \bar{x}(2)] - [z - \bar{x}(1)]' S_1^{-1} [z - \bar{x}(1)] \geq \log \frac{|S_1|}{|S_2|}.$$

ut these quantities were evaluated numerically above. The left-hand de is $\frac{4}{7} - \frac{1}{3}$. Hence the rule is to classify z into π_1 if

$$\log \frac{|S_1|}{|S_2|} \leq \frac{4}{7} - \frac{1}{3} = \frac{5}{21} = .238.$$

ut it was also found above that $|S_1| = \frac{3}{4}$, and $|S_2| = \frac{7}{16}$. Hence,

$$\log \frac{|S_1|}{|S_2|} = \log \frac{12}{7} = \log 1.71 = .536.$$

Therefore, z should be classified into π_2, the same conclusion reached by the Bayesian approach. However, in this case there is great uncertainty about the decision since large sample theory was used (and there was no reason to suppose it was valid to do so). Moreover, while the Bayesian approach provided a complete predictive distribution (or a continuum of "risks") for placing z in π_2 (in addition to the predictive odds), the sampling theory result provided only a decision. Sampling theory procedures that attempt to cope with the problem of risk associated with the classification decision in the absence of well-defined loss functions still fall short in that it is then required that sample sizes be large, and that the covariance matrices be equal, assumptions that are not always justified.

13.4 TEST FOR DISCRIMINATORY POWER

After a discrimination procedure has been established, it is of considerable interest to determine whether the discriminator is really useful. A method for studying the discriminatory power of a procedure involves the use of *Confusion matrices*, which were defined by Massy (1965) for comparing the similarities among populations.

A Confusion matrix provides a convenient method of summarizing the number of correct and incorrect classifications made by the discrimination procedure. Suppose there are K populations and N_j observations have been taken from π_j to estimate its parameters, $j = 1, \ldots, K$. Since the origins of all these observations are known, by applying the discrimination procedure to these observations, it is possible to score the fraction of successful classifications, and to test whether the procedure is significantly better than a purely random partitioning of the decision space.

Let n_{ij} denote the number of observations known to belong to population π_i, but which were classified into π_j. Then the Confusion matrix for the classification problem is defined to be the $K \times K$ matrix $C \equiv (n_{ij})$ depicted below.

$$\underset{(K \times K)}{C} = \begin{array}{c|cccc|} & \pi_1 & \pi_2 & \cdots & \pi_K \\ \hline \pi_1 & n_{11} & n_{12} & \cdots & n_{1K} \\ \pi_2 & n_{21} & n_{22} & \cdots & n_{2K} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \pi_K & n_{K1} & n_{K2} & \cdots & n_{KK} \\ \hline \end{array} \quad \text{True } \pi_i\text{'s}$$

Confusion matrix — Predicted π_j's

onal elements of C denote the numbers of correct classifications), and the off-diagonal elements denote the numbers of incorrect ifications (misses). The *normalized Confusion matrix*, $\mathbf{C}_0$, is easier to pret than C. By definition,

$$\mathbf{C}_0 = (c_{ij}), \qquad c_{ij} = \frac{n_{ij}}{\sum_{j=1}^{K} n_{ij}}.$$

t is, the elements of the normalized Confusion matrix are fractions of ect and incorrect classifications.

o test the discriminatory power of the procedure use a chi-square Accordingly, define

$$Q = \frac{(n-e)^2}{e} + \frac{(\bar{n}-\bar{e})^2}{\bar{e}}, \tag{13.4.1}$$

re n and $\bar{n}$ denote the number of correct and incorrect classifications le by the discrimination procedure, respectively, and e and $\bar{e}$ denote expected numbers of correct and incorrect classifications that would nade if the classifications were made at random. Then, if $N \equiv \Sigma_1^K N_j$ otes the total number of observations classified, and the probability successful random classification is $1/K$,

$$n = \sum_{j=1}^{K} n_{jj}, \qquad \bar{n} = N - n, \qquad e = \frac{N}{K}, \qquad \bar{e} = N - \frac{N}{K}. \tag{13.4.2}$$

s easy to check by substituting the relations in (13.4.2) into (13.4.1) t the test statistic is expressible as

$$Q = \frac{(N - nK)^2}{N(K-1)}, \tag{13.4.3}$$

orm more convenient for numerical evaluation. Thus, to test H: hits k place at random versus A: the discrimination procedure did better n just chance, use the fact that under H,

$$\mathcal{L}(Q) = \chi_1^2. \tag{13.4.4}$$

t should be noted that since the same data is being used to rate the cedure as to define the procedure, the test for discriminatory power not strictly appropriate. A correct test would be obtained by splitting sample into one part which is used to establish the discrimination procedure and another which is used to test the procedure. However, if the sample is small, this approach is not recommended since then, inefficient estimators of the parameters would result at the expense of obtaining a good power testing procedure.

Example (13.4.1): Consider Example (13.3.1) in which audiences of 5 FM radio stations in Boston were classified according to 12 socioeconomic characteristics. The Confusion matrix and the normalized Confusion matrix for this problem are given in Tables 13.4.1 and 13.4.2. Results are based upon 239 observations with known classifications. Adding the diagonal elements of Table 13.4.1 shows that the total number of hits was 88, or 36.8 percent. Evaluating Q from (13.4.3) for $N = 239$, $n = 88$, and $K = 5$ gives $Q = 42.2$. Since at the 1 percent level, $\chi_1^2 = 10.8$, Q is certainly significant, and H must be rejected. Thus, the classification procedure does better than chance.

Table 13.4.1 Confusion Matrix for Radio Audiences

Actual audience	*Predicted audience*					*Totals*
	A	*B*	*C*	*D*	*E*	
A	43	13	8	21	14	99
B	16	15	15	13	13	72
C	3	5	14	5	4	31
D	2	3	5	9	4	23
E	2	1	0	4	7	14

Table 13.4.2 Normalized Confusion Matrix for Radio Audiences

Actual audience	*Predicted audience*				
	A	*B*	*C*	*D*	*E*
A	.43	.13	.08	.21	.14
B	.22	.21	.21	.18	.18
C	.10	.16	.45	.16	.13
D	.08	.13	.22	.39	.17
E	.14	.07	.00	.29	.50

Other examples of the use of discrimination analysis in marketing and business were given by Banks (1958), Evans (1959), and Frank and Massy (1963).